Київський національний університет імені Тараса Шевченка
механіко-математичний факультет

ПРОЄКТНА РОБОТА
зі статистики на тему
«Прогнозування переможця премії „Оскар“ в номінації „Найкращий фільм року“ в 2021 році»

Студентів 3-го курсу
спеціальності „111 Математика“
Вербицької Олени
Мацуй Анни
Мальцевої Діани
Аніщенко Романа
Бурого Тараса

Науковий керівник:
професор, доктор фіз.-мат. наук
Шевченко Георгій Михайлович

Київ — 2021

Зміст

1 Вступ

Сьогодні формально існує три підходи для прогнозування та аналізу часових рядів:

- економетричний, регресійний;
- методи Бокса-Дженкінса (ARMA, ARIMA);
- методи, які засновані на використанні штучної нейронної мережі.

В цій роботі ми хочемо продемонструвати, як працюють різні методи прогнозування для часових рядів, які описують фільми-переможці в номінації „Найкращий фільм року“ премії „Оскар“. Для цього ми будемо використовувати п'ять часових рядів, а саме: „Age“ – вік актора, що грає головну роль, „Minor“ – кількість акторів, що відносяться до меншин, „W/M“ – відношення жінок до чоловіків серед акторів, „Nomin“ – кількість номінацій фільму в році отримання „Оскара“ і „Box office“ – касові збори фільму. Значення всіх цих часових рядів знаходяться в датасеті, який ми збирали власноруч, за посиланням https://docs1.google.com

В підсекціях «Прогнозування за допомогою наївних методів» та «Прогнозування за допомогою експоненційного згладжування» представлені найпростіші економетричні методи та пояснено, чому деякі з них не підходять для опису цих п'яти часових рядів. Також в підсекції «Прогноз часового ряду „Box office“ за допомогою апроксимації алгебраїчними поліномами» наведений ще один економетричний метод прогнозування для часового ряду „Box office“ та описані недоліки цього методу.

Секція «Модель ARMA(p, q) для часового ряду „Box office“» повністю присвячена методу Бокса-Дженкінса, проте аналіз та прогноз робиться тільки для одного часового ряду „Box office“. В підсекції «Чи потрібно позбавлятися від викидів?» описано, чому саме цей ряд було обрано.

Застосувати методи, які засновані на використанні нейронної мережі, нам так і не вдалося, адже вони потребують великих об'ємів тренувальних та тестових даних, тому натренувати алгоритм на дуже обме-

женій вибірці з 92 років виявилося неможливим. За допомогою машинного навчання (логістичної регресії) був знайдений фільм-переможець. При цьому використано датасет з рейтингами усіх фільмів, що номінувались на „Оскар“, починаючи з 1944 року. Посилання на датасет https://docs2.google.com

2 Прогнозування за допомогою логістичної регресії

2.1 Як працює логістична регресія

Нехай є деяка випадкова величина Y, що може набувати лише двох значень, які, як правило, позначаються цифрами 0 і 1. Нехай ця величина залежить від деякої множини пояснювальних змінних $x = (1, x_1, \ldots, x_n)^T$. Залежність Y від $x_1, \ldots, x_n$ можна визначити, ввівши додаткову змінну y^*, де

$$y^* = \theta^T x = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n + \varepsilon.$$

Тоді:

$$Y = \begin{cases} 0, \ y^* \leq 0, \\ 1, \ y^* > 0 \end{cases}$$

При визначенні логістичної моделі стохастичний доданок ε вважається випадковою величиною з логістичним розподілом ймовірностей. Логістичний розподіл — неперервний ймовірнісний розподіл, що за формою нагадує нормальний розподіл, проте має більший коефіцієнт ексцесу. Функція щільності логістичного розподілу визначається за формулою

$$f(x; \mu, s) = \frac{e^{-(x-\mu)/s}}{s\left(1 + e^{-(x-\mu)/s}\right)^2}.$$

Для певних конкретних значень змінних $x^* = x_1^*, \ldots, x_n^*$ одержується відповідне значення y^* і ймовірність того, що $Y = 1$ така:

$$p(Y = 1) = p(y^* > 0) = p(\theta^T x^* + \varepsilon > 0) = \\ p(\varepsilon > -\theta^T x^*) = p(\varepsilon \leqslant \theta^T x^*) = \Lambda(\theta^T x^*)$$

Передостання рівність випливає з симетричності логістичного розподілу, а Λ позначає логістичну функцію — функцію розподілу логістичного розподілу:

$$\Lambda(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Таким чином, для конкретного значення x^i випадкова величина Y^i має розподіл Бернуллі: $Y^i \sim B(1, \Lambda(\theta^T x^i))$.

Логістична регресія задовольняє наступну умову:

$$\ln \frac{p(1|X)}{1-p(1|X)} = \ln \frac{p(1|X)}{p(0|X)} = b_0 + b_1 x_1 + ... + b_J x_J$$

Оцінка параметрів $\theta_0, \theta_1, ..., \theta_n$, яка робиться на основі деякої вибірки $(x^{(1)}, Y^{(1)}), ..., (x^{(m)}, Y^{(m)})$, де $x^{(i)} \in \mathbb{R}^n$ — вектор значень незалежних змінних, а $Y^{(i)} \in \{0, 1\}$ — відповідне їм значення Y, здійснюється за допомогою методу максимальної вірогідності, згідно з яким вибираються параметри θ, що максимізують значення функції вірогідності на вибірці:

$$\hat{\theta} = \text{argmax}_\theta L(\theta) = \text{argmax}_\theta \prod_{i=1}^{m} \Pr\{Y = Y^{(i)} | x = x^{(i)}\}$$

2.2 Результати прогнозування

За критерії, що можуть спрогнозувати переможця, було взято рейтинги фільмів з різних сайтів - RottenTomatoes (оцінка як і від критиків, так і від глядачів), IMDB, MetaScore.

Оскільки протягом років кількість фільмів-номінантів була різною, вважаємо, що щороку номінувалось 10 фільмів, додаючи фільми з рейтингами 0 до відповідних років. Таким чином, вхідні дані за один рік - вектор розмірності (1, 40), оскільки є 4 рейтинги для кожного з 10 фільмів. Звідси маємо отримати вектор розмірності (1, 10), де для кожного фільму шукається імовірність його перемоги. Очевидно, для всіх років, окрім 2021, в цьому векторі будуть значення 0 для всіх, окрім вже відомого переможця, у якого буде значення 1.

Цього року номінувалось 7 фільмів. А саме:

- „Nomadland“;
- „The Father“;
- „Judas and the Black Messiah“;
- „Mank“;
- „Minari“;
- „Promising Young Woman“;
- „Sound of Metal“;
- „The Trial of the Chicago 7“.

За допомогою логістичної регресії було отримано, що цього року мав би перемогти „The Trial of the Chicago 7“. Легко переконатись, що насправді переміг інший фільм, проте сайти та статті критиків ставили цей фільм на друге місце. Також поза оцінками критиків враховуються різні інші мотиви, які тут не враховувались. Тренувальною вибіркою вважалися всі роки, починаючи з 1980, оскільки для старіших фільмів рейтинги могли змінитись з часом.

3 Основні відомості про часові ряди

Часовий ряд (time series) – це множина значень y_1, y_2, $\ldots$, y_t, отриманих в результаті спостережень за певним явищем в рівні проміжки часу $[t_1; t_2] = [t_2; t_3] = \cdots = [t_{n-1}; t_n]$. Якщо час змінюється дискретно, то часовий ряд називається дискретним. Ми будемо розглядати тільки дискретні часові ряди, в яких спостереження записуються через фіксований інтервал часу.

Іноді нам недостатньо тільки спостерігати певні процеси або знаходити закономірність між елементами різних статистичних вибірок, тому однією з основних задач аналізу часових рядів є так звана задача прогнозування, що полягає у знаходженні такої функції f_t, що

$$y_{t+h} \approx f_t(y_1, y_2, \ldots, y_t, h) \equiv \hat{y}_{t+h},$$

де $h \in \{1, 2, \ldots, H\}$, H – яку кількість майбутніх значень ми хочемо дізнатися.

При прогнозувані часових рядів слід враховувати, що

- аналізуються лише дані спостережень без додаткової інформації та без аналізу впливу зовнішніх сил. Звичайно, такий аналіз виглядає досить неповним, але доволі часто прогнози часових рядів є більш точними;

- в багатьох задачах статистики робиться припущення щодо незалежності спостережень y_1, y_2, $\ldots$, y_t, проте в прогнозувані часових рядів, навпаки, значення y_t вважаються залежними від попередніх значень y_{t-1}, y_{t-2}, $\ldots$, y_1. Чим більше зв'язків та закономірностей між елементами часового ряду нам вдасться знайти, тим краще буде прогноз функцію f_t.

Кожен часовий ряд y_t, для якого розв'язується задача прогнозування, є сумою двох складових, а саме:

$$y_t = l_t + \varepsilon_t,$$

де l_t – систематична складова, яку ми намагаємося виявити та описати статистичним методами, ε_t – випадковий шум, який ускладнює виявлення регулярних (систематичних) компонент.

Розглянемо, які компоненти входять до систематичної складової l_t часового ряду y_t.

- Тренд $T(t)$ – лінійна або нелінійна систематична компонента, яка повільно змінюється з плином часу.

- Сезонність $S(t)$ – це компонента, яка періодично повторюється.

- Цикл $C(t)$ – компонента, яка відповідає за зміну рівня ряду зі зміною періоду.

Наприклад на графіку **Рис.1** можна помітити наявність двох систематичних компонент, а саме тренд (червона лінія), оскільки середнє зна-

чення часового ряду повільно змінюється з плином часу, та сезонність, оскільки кожен рік, графік майже повністю повторюється.

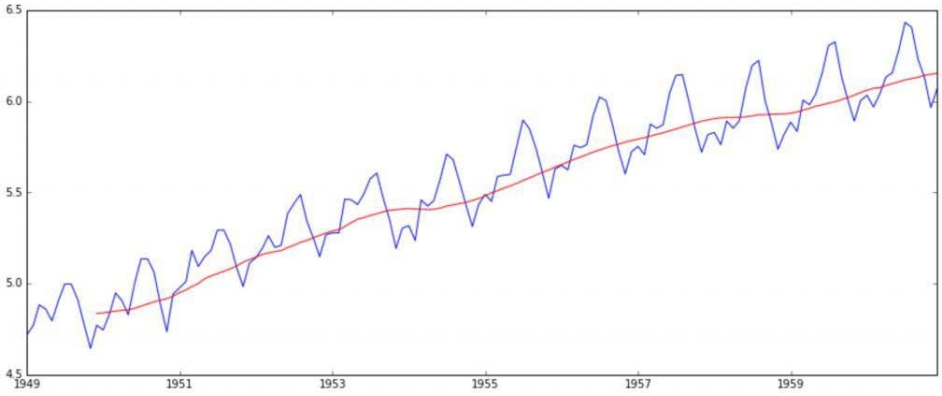

Рис. 1: Часовий ряд з трендом та сезонністю

Циклічну компоненту легко можна побачити на графіку **Рис.2**. Цей часовий ряд має деяку сезонну структуру, проте кожен рік з'являються якість нові піки, та і взагалі часовий ряд стає гострішим.

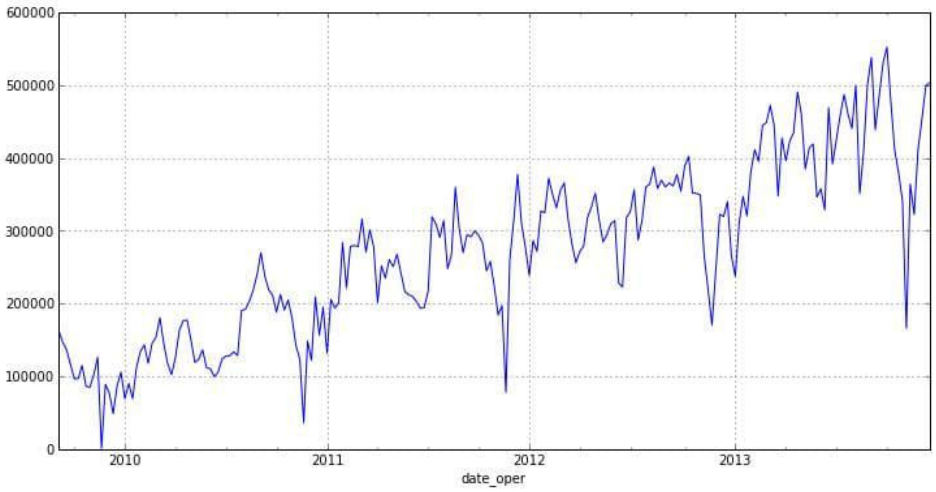

Рис. 2: Часовий ряд з циклом

В подальшому аналізі нам ще знадобиться визначення одного спеціального класу часових рядів. Часовий ряд називається *стаціонарним*, якщо його властивості не залежать від часу, тобто:

- $\mathbb{E}(y_1) = \mathbb{E}(y_2) = \mathbb{E}(y_3) = \cdots = \mu$,
- $\mathbb{D}(y_1) = \mathbb{D}(y_2) = \mathbb{D}(y_3) = \cdots = \sigma^2$,

- $Cov(y_1, y_2) = Cov(y_2, y_3) = Cov(y_3, y_4) = \cdots = \gamma$,

де μ, σ^2 и γ математичне сподівання, дисперсія та коефіцієнт коваріації відповідно.

4 Економетриичні методи прогнозування

4.1 Прогнозування за допомогою наївних методів

Іноді нам потрібно швидко знайти майбутнє значення деякого часового ряду y_t, тобто просто зробити грубу оцінку значень y_{t+1} y_{t+1}, $\ldots$, y_{t+H} не шукаючи залежності між елементами часового ряду. Для цього в статистичному аналізі існують так звані наївні моделі.

При створенні наївних моделей передбачається, що деякий останній період прогнозованого тимчасового ряду найкраще описує майбутнє цього прогнозованого ряду, тому в цих моделях прогноз, як правило, є дуже простою функцією, що залежить від значень прогнозованої змінної в недалекому минулому.

В цій роботі ми спрогнозуємо п'ять часових рядів за допомогою трьох наївних моделей.

Першу модель назвемо „завтра буде те саме, що було сьогодні“. Функція прогнозу даної моделі описується за допомогою формули

$$y_{t+1} = y_t,$$

де y_t – останнє відоме значення ряду, y_{t+1} – наш прогноз на один рік вперед.

Пісня застосування цієї моделі, ми отримаємо такі значення „прогнозу“ для відповідних часових рядів

Name	Age	Minor	W/M	Nomin	Box office
y_{2021}	52	0	0.625	6	254.1

Звісно, від такої примітивної моделі не варто чекати великої точності. Вона не тільки не враховує механізми, що визначають прогнозовані дані

(цей серйозний недолік взагалі властивий багатьом статистичним методам прогнозування), але і не захищена від випадкових флуктуацій, вона не враховує сезонні коливання, цикли та тренди. Втім, можна будувати наївні моделі дещо по-іншому, а саме

$$y_{t+1} = y_t + (y_t - y_{t-1}).$$

Цей прогноз хоча і є дуже схожим на попередній, проте вже враховує зміни, які відбулись між значенням y_{t-1} та y_t. Тому можна вважати, що ця наївна модель частково враховує трендову компоненту часового ряду. Прогнози цієї моделі зображені в таблиці

Name	Age	Minor	W/M	Nomin	Box office
y_{2021}	44	-2	1.125	7	179.6

Бачимо, що цей прогноз може навіть дати значення, які не можуть існувати, оскільки кількість меншин ("Minor") дорівнює -2, що, звичайно, неможливо.

Останній наївний метод – це метод відшукання середнього арифметичного. Формула для прогнозу виглядає наступним чином

$$y_{t+1} = \frac{y_t + y_{t-1} + \cdots + y_{t-k}}{k},$$

де k – це кількість попередніх значень, за якими береться середнє арифметичне. Найчастіше k обирають з множини чисел $\{3, 5, 7\}$. Таблиця прогнозів виглядає таким чином

Name	Age	Minor	W/M	Nomin	Box office
y_{2021}, $k = 3$	51	1	0.39	8	259.3
y_{2021}, $k = 5$	45.4	2.8	0.32	7.6	188.3
y_{2021}, $k = 7$	46.57	5.8	0.48	8	176.05

Може здаватися, що будувати такі моделі досить безглуздо, проте це, звичайно, не так. Найчастіше їх використовують, щоб тестувати більш складні методи. Продемонструємо, як застосовуються ці моделі в наступній підсекції.

4.2 Прогнозування за допомогою експоненційного згладжування

Як було зазначено вище, існує багато різних методів прогнозування часових рядів. Деякі з них (наприклад, наївні методи), не передбачають перевірки часових рядів на стаціонарність, наявність трендових, сезонних та циклічних компонент. Інші же, навпаки, потребують попереднього статистичного аналізу та більш глибокого розуміння теорії. На перший погляд може здаватися, що більш складні методи прогнозування дають кращий прогноз, ніж елементарні методи, проте це не завжди так. Саме тому результати моделі прогнозування часто порівнюють з результатами прогнозів декількох наївних моделей, оскільки якщо прогноз однієї з наївних моделей виходе кращим, ніж при використанні більш складного методу, то це говорить лише про те, що обраний складний метод не підходить для прогнозування цього часового ряду.

Продемонструємо це явище на прикладі використання методу експоненційного згладжування (Simple Exponential Smoothing, SES). Експоненціальне згладжування здійснюється за допомогою рекурентної формули

$$s_t = \begin{cases} y_0 & , t = 0, \\ s_{t-1} + \alpha(y_t - s_{t-1}) & , t \neq 0, \end{cases}$$

де s_t – значення згладженого ряду, y_t – значення початкового ряду та α – коефіцієнт згладжування, для якого виконується $\alpha \in (0, 1)$. Як і будь-яке згладжування, метод SES допомагає зменшити вплив випадкової компоненти ε_t, яку, звісно, неможливо спрогнозувати, в часовому ряді $y_t = l_t + \varepsilon_t$, де l_t – компонента часового ряду, яку можна видалити, використовуючи статистичні методи. Дійсно, якщо дисперсію початкового ряду покласти рівною σ^2, то після експоненційного згладжування її значення зменшиться до $\sigma^2\alpha/(2-\alpha)$, тобто зі зменшенням коефіцієнту α вплив випадкової компоненти ε_t стає менш суттєвим. Тому краще обирати коефіцієнт згладжування α близький до 0. Проте якщо більш детально розглянути рекурентну формулу для SES, то можно помітити,

що зі збільшенням коефіцієнта α збільшується доданок αy_t. Це говорить про те, що при прогнозуванні модель краще реагує на останні значення часового ряду y_t, y_{t-1}, ..., а початкові значення y_0, y_1, ... практично не беруть участь в прогнозуванні, що, звичайно, нам вигідно, оскільки майже всі значення часових рядів, наприклад, в 1930-их роках, сильно відрізняються від значень в 2000-их роках. Таким чином, ми отримали дві протилежні умови на α: з одного боку потрібно, щоб $\alpha \to 0$, а з іншого, щоб $\alpha \to 1$.

На практиці коефіцієнт згладжування α часто шукається за допомогою методу сітки. Можливі значення параметра розбивають сіткою з деяким кроком. Коефіцієнт α підбирають таким чином, щоб середнє від квадрату відхилення (Mean Squared Error, MSE) згладжуваного ряду s_t від початкового ряду y_t було мінімальне, тобто

$$\{\ \alpha \in (0,1) \mid \min \mathrm{MSE} = \min \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=0}^{\mathrm{n}} (\mathrm{y_t} - \mathrm{s_t})^2\ \}.$$

Спробуємо підібрати оптимальний коефіцієнт згладжування використовуючи $\alpha \in \{0.1,\ 0.3,\ 0.5,\ 0.6,\ 0.7,\ 0.9,\ 0.95\}$ для часових рядів „Age“, „Minor“, „W/M“, „Nomin“ і „Box office“. В таблиці нижче можна побачити значення MSE для отриманих згладжених часових рядів при відповідних значеннях α.

α	0.1	0.3	0.5	0.6	0.7	0.9	0.95
Age	89.83	55.91	32.2	22.1	13.37	1.73	0.45
Minor	17.13	11.3	6.65	4.61	2.82	0.37	0.097
W/M	0.41	0.28	0.16	0.11	0.07	0.01	0.002
Nomin	6.63	3.25	1.72	1.15	0.68	0.088	0.022
Box office	51019.33	31922.29	18701.01	12979.67	7956.32	1063.68	279.281

З вигляду таблиці можна зробити висновок, що MSE спадає при збільшені α, причому мінімальне значення MSE не вдалося знайти на проміжку $(0,1)$, бо коли ми будемо розглядати значення $\alpha \to 1$, то MSE$\to 0$. Покладемо $\alpha = 1$, тоді з рекурентної формули для експоненційного згладжування випливає, що прогноз $s_{2021} = y_{2020}$, тобто прогноз на 2021 рік

обраних нами часових рядів співпадає зі значенням цих часових рядів в 2020 році. Цей прогноз співпадає з наївним прогнозом моделі „завтра буде те саме, що було сьогодні“. Таким чином модель SES не дає кращі прогнози для обраних параметрів, порівняно з однією з наївних моделей, тому SES не підходить для нашого аналізу.

5 Попередній аналіз часового ряду „Box office“

Надалі ми будемо займатись аналізом лише одного часового ряду – касових зборів фільмів-переможців (колонка „Box office“). Саме цей часовий ряд було обрано у зв'язку з наявністю у його графіку певних закономірностей, які можна помітити без попереднього статистичного аналізу.

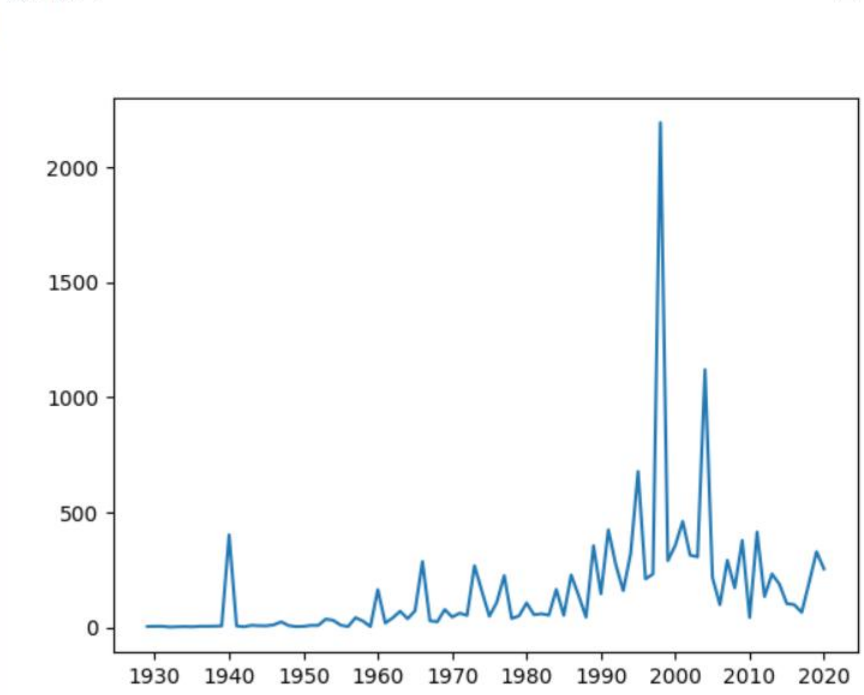

Рис. 3: Часовий ряд „Box office“

В цьому графіку присутня трендова компонента, що відповідає за плавні зміни, а після 1959 року піки змінюються падіннями щороку або кожні два роки (сезона компонента). Також відмітимо зростання дисперсії з плином часу.

5.1 Чи потрібно позбавлятися від викидів?

Зазвичай перший етапом знаходження в будь-якій вибірці закономірностей, таких як опис кривої росту чи знаходження функції розподілу, є аналіз вибірки на наявність так званих викидів. Після того, як у вибірці вдається знайти аномальне значення, від нього позбавляються або ж використовують різноманітні способи згладжування. Відмінність нашої системи від кратної вибірки полягає в тому, що значення часового ряду y_t в момент часу t вважається залежним від попередніх значень, а саме:

$$y_t \approx f(y_{t-1}, y_{t-2}, \ldots, y_1).$$

Через це згладжування викидів може призвести до втрати певної закономірності між значеннями часового ряду та погіршити якість прогнозу. За допомогою автокореляції продемонструємо зміну структури часового ряду „Box office“ при згладжування викидів. Як відомо, кореляція характеризує степінь статистичної взаємодії між елементами даних. Відміна автокореляції k-го порядку від звичайної кореляції полягає лише у тому, що степінь залежності визначається між значеннями одного й того самого часового ряду. При цьому обчислюються значення кореляції між початковим часовим рядом та його копією, що зміщена на k значень (так званий лаг k-го порядку). Загальна формула для знаходження автокореляції r_k при лазі k-го порядку виглядає наступним чином

$$r_k = \frac{\mathbb{E}((y_t - \mathbb{E}y)(y_{t+k} - \mathbb{E}y))}{\mathbb{D}y}, \qquad r_k \in [-1; 1].$$

Наприклад, вибіркова автокореляція r_1 першого порядку обчислюється за формулою

$$r_1 = \frac{\sum_{i=2}^{n}(y_i - \bar{y}_1)(y_{i-1} - \bar{y}_2)}{\sqrt{\sum_{i=2}^{n}(y_i - \bar{y}_1)^2 \sum_{i=2}^{n}(y_{i-1} - \bar{y}_2)^2}}, \quad \bar{y}_1 = \frac{\sum_{i=2}^{n} y_i}{n-1}, \quad \bar{y}_2 = \frac{\sum_{i=2}^{n} y_{i-1}}{n-1}.$$

Для часового ряду, зображеного на рис. 1, таблиця лагів до 69-го порядку може бути представлена у вигляді

За даними цієї таблиці може бути порахована вибіркова автокореляція та побудована автокореляційна функція (АКФ) – функція залежності коефіцієнту автокореляції від кількості лагів.

y_i	3.6	4.4	4.6	2.59	1.38	2.59	...	65	195.2	328.6	254.1
y_{i-1}		3.6	4.4	4.6	2.59	1.38	2.59	...	65	195.2	328.6
y_{i-2}			3.6	4.4	4.6	2.59	1.38	2.59	...	65	195.2
...											
y_{i-69}											3.6

Значення АКФ для ряду „Box office“

k	0	1	2	3	4	5	...	66	67	68	69
r_k	1	0.21	0.22	0.42	0.25	0.2	...	-0.14	-0.038	-0.095	-0.007

та її графік зображені на **Рис. 4**. Дивлячись на цей графік можна

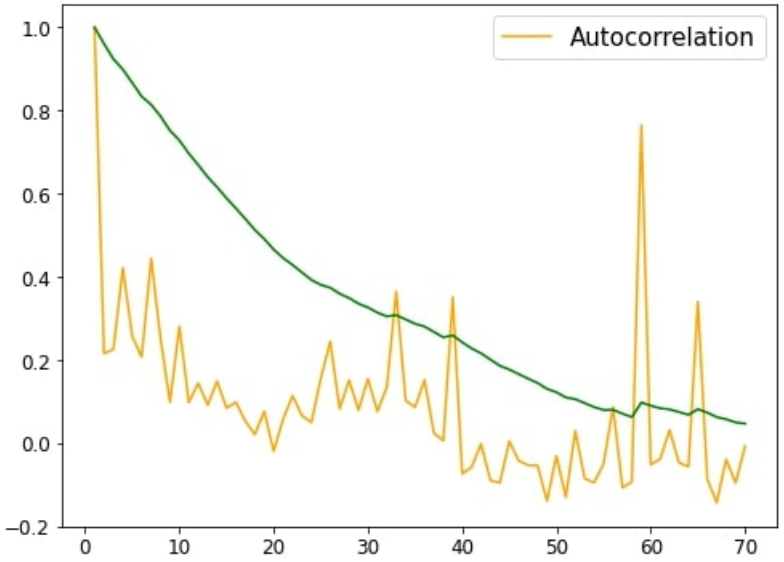

Рис. 4: Функція автокореляції для часового ряду „Box office“

провести деякий неформальний аналіз, оцінити, які компоненти (тренд, сезонність, циклічність) присутні в часовому ряді. На **Рис. 4** можна одразу побачити, що значення зелена крива демонструє спадання коефіцієнта автокореляції при збільшені порядку лагу, що може говорити про наявність тренду. Приблизно кожні двадцять лагів автокореляція майже не відрізняється від нуля $r_{20} = -0.019$, $r_{42} = -0.001$, $r_{62} = 0.031$, що свідчить про наявність циклів, а зміну максимумів мінімумами можна інтерпретувати, як вплив сезонності на часовий ряд. Таким чином, ми

"виявили"[1] одразу три структури у цього часового ряду, що може бути корисним при його прогнозуванні.

Тепер спробуємо знайти викиди та застосувати один з способів згладжування. Для виявлення аномальних значень ми використали критерій Ірвіна, згідно з яким аномальним вважається значення y_t, що відрізняється від попереднього значення y_{t-1} на величину, більшу ніж середньоквадратичне відхилення

$$\lambda_i = \frac{|y_t - y_{t-1}|}{\sigma},$$

де λ_i – критерій Ірвіна, σ – середньоквадратичне відхилення. Значення вважається викидом, якщо $\lambda_i > \lambda(n)_{tab}$ (в нашому випадку $\lambda(n = 92)_{tab} = 1$). Цей критерій виявив вісім аномальних значень, три з яких $y_{12} = 402$, $y_{70} = 2194.4$ та $y_{76} = 1120.4$ можна легко помітити на рис.1. Згладжування восьми викидів проводилось за допомогою середнього по трьом, п'ятьом та по семи точкам. Кількість елементів при згладжуванні визначалась так, щоб зберігались основні співвідношення між значеннями числового ряду. Таким чином, отримуємо числовий ряд **Рис. 5** та графік АКФ **Рис. 6** для цього часового ряду.

З вигляду графіка АКФ можна також припустити, що отриманий часовий ряд має трендову та сезонну компоненти, проте вже важко говорити про існування циклічної компоненти, оскільки відсутні приблизні повтори графіку АКФ через кожні двадцять лагів.

Таким чином, змінивши значення всього восьми точок часового ряду, ми, можливо, втратили суттєву інформацію про початковий ряд. Тому надалі ми будемо розглядати задачу прогнозування лише початкового ряду „Box office".

[1]Звичайно, наявність тих чи інших компонент в часовому ряді потрібно перевіряти статистичними методами, проте ми не будемо на цьому зосереджуватися.

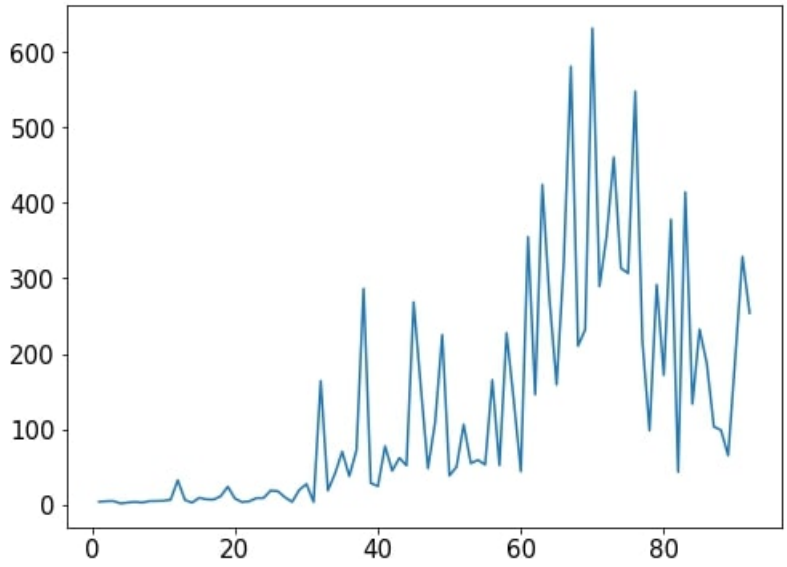

Рис. 5: Графік часовий ряд „Box office“ зі згладженими викидами

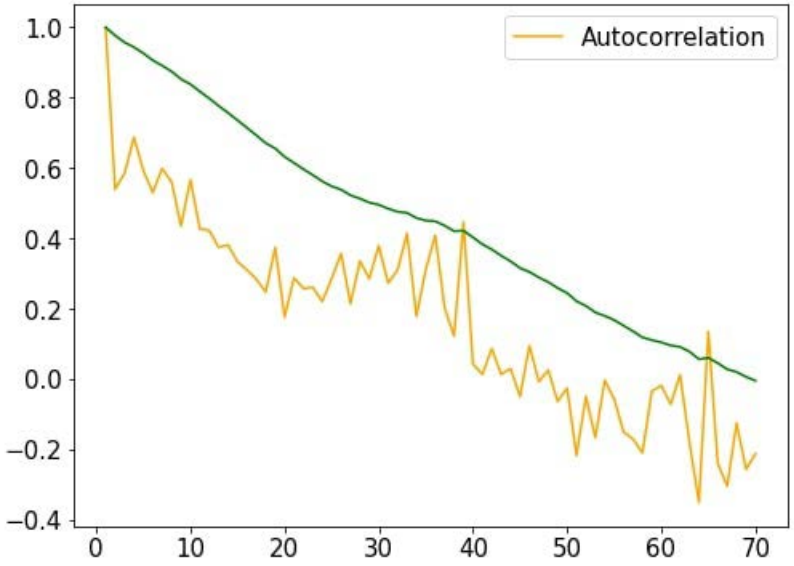

Рис. 6: Функція автокореляції для згладженого часового ряду „Box office“

5.2 Прогноз часового ряду „Box office“ за допомогою апроксимації алгебраїчними поліномами

Метод найменших квадратів (МНК) застосовують, коли потрібно вивести формулу апроксимуючої кривої, що описує деяку залежність, отриману в результаті експеримента. Оскільки експериментальні дані отримують, як правило, з деякою похибкою, то немає сенсу використовувати, наприклад, інтерполяційний поліном Лагранжа, що завжди проходить

через вузли інтерполяції. В такому випадку зазвичай проводять апроксимуючу криву, що не проходить через експериментальні точки, але враховує досліджувану закономірність, згладжуючи викиди результатів експеримента.

За допомогою функції polyfit, яка використовує МНК, в Python було отримано наближення поліномом $p(x)$ 7-го степеня для часового ряду „Box office“ **Рис.7**.

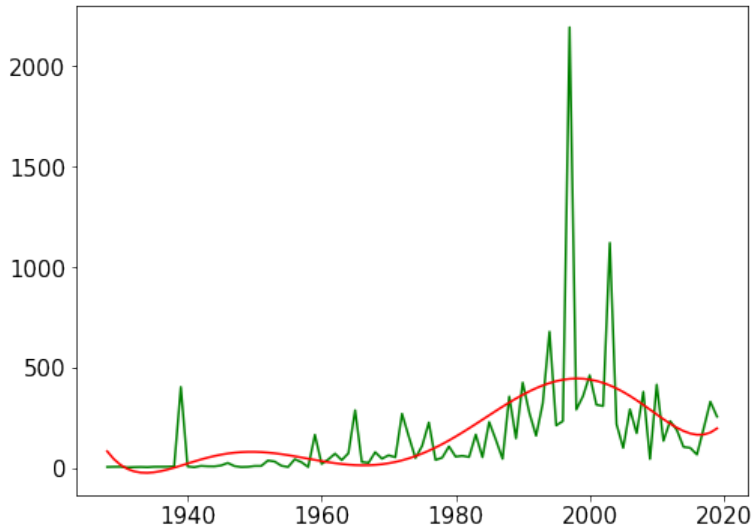

Рис. 7: Наближення часового ряду „Box office“ алгебраїчним поліномом 7-го степеня

Розглянемо залишки r_t, $t \in \{1928, 1929, \ldots, 2020\}$ – різниці реальних даних та наближених поліномом в момент часу t

$$r_t = y_t - p(t)$$

Дослідимо їх розподіл на нормальність. Можна зробити це графічним методом за допомогою гістограми та графіка Q-Q (quantile-quantile plot) – графіка, що покаже залежність між теоретичними та вибірковими квантилями. Якщо ця залежність лінійна, то вибірка нормально розподілена.

Як бачимо, гістограма розподілу **Рис.8** несиметрична, а точки на графіку Q-Q **Рис.9** не лежать на одній прямій, тому ми можемо припустити, що розподіл залишків не є нормальним.

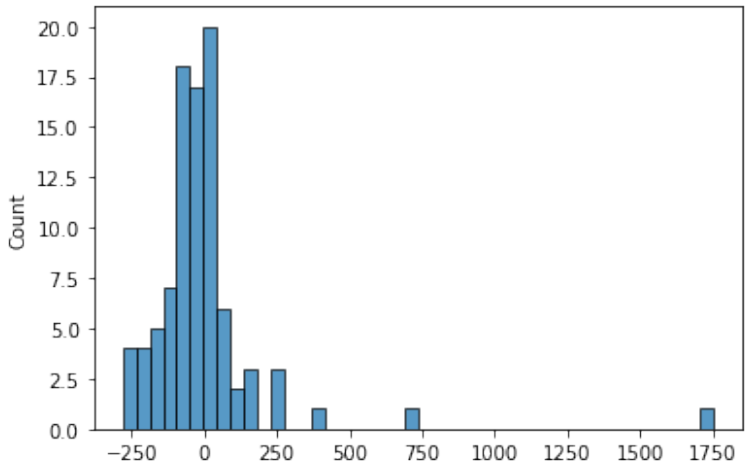

Рис. 8: Гістограма залишків r_t

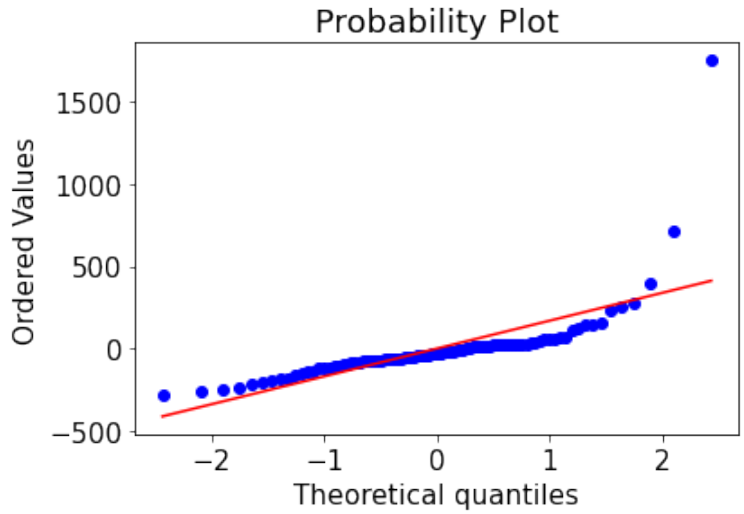

Рис. 9: Q-Q графік залишків r_t

Більш точним тестом на нормальність є тест Шапіро-Уілка (Shapiro-Wilk normality test), що також є придатним і для таких невеликих вибірок, як залишки r_t. За допомогою цього тесту було підтверджено наше припущення про ненормальність розподілу.

З цього можна зробити висновок, що в залишках ще залишилась певна залежність. З графіка залишків **Рис.10** можна помітити деяку циклічність, що свідчить про те, що апроксимація поліномами погано наближує реальний часовий ряд, залишаючи багато інформації про початковий ряд в залишках.

Для порівняння на рисунку **Рис.11** зображено нормально розподілені

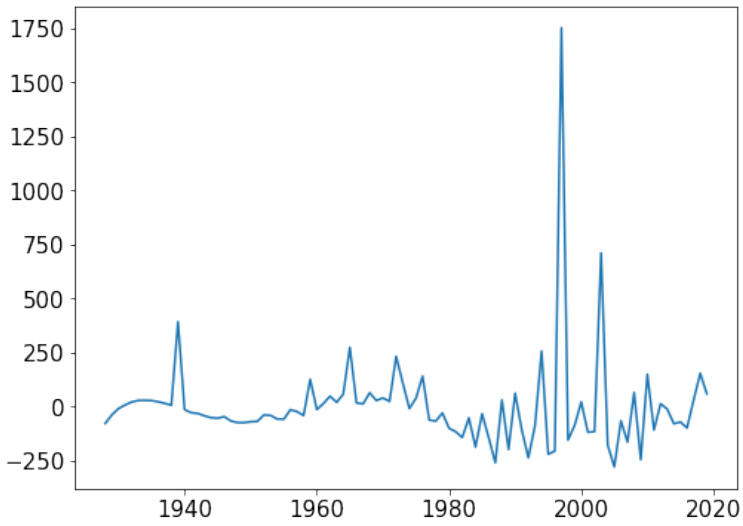

Рис. 10: Графік залишків r_t

залишки – білий шум.

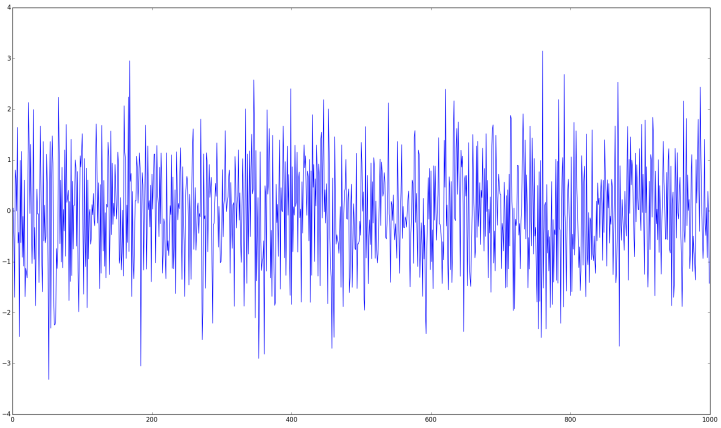

Рис. 11: Графік білого шуму з нормальним розподілом $N(0, \sigma^2)$

6 Модель ARMA(p, q) для часового ряду „Box office“

Деякі моделі в аналізі часових рядів займають особливе місце та частіше за інші моделі використовуються на практиці. Це пов'язано з наявністю у цих моделях цікавих властивостей, які дозволяють не тільки описати початковий часовий ряд та знайти прогноз на декілька кроків, а ще і побудувати довірчий інтервал для цього прогнозу та перевірити часовий

ряд на стаціонарність. Саме ці моделі ми будемо використовувати в цій роботі.

6.1 Загальна теорія про AR(p), MA(q) та ARMA(p, q) моделі

Авторегресійною моделлю порядку p (AR(p)) називають модель часового ряду, в якій значення часового ряду в момент часу y_t лінійно залежить від попередніх значень того ж самого ряду. Цю модель представляють у вигляді формули

$$y_t = c + \sum_{i=1}^{p} a_i y_{t-i} + \varepsilon_t,$$

де c – деяка стала, p – кількість попередніх значень y_{t-1}, y_{t-2}, …, y_{t-p}, ε – білий шум з $\mathbb{E}(\varepsilon_t) = 0$ та $\mathbb{D}(\varepsilon_t) = \sigma_\varepsilon$, a_i – деякі сталі. Цю формулу також зручно записати, використовуючи так званий оператор лагу $L^i := L^i y_t = y_{t-i}$, а саме:

$$y_t = (\sum_{i=1}^{p} a_i L^i) y_t + \varepsilon_t \qquad \text{або} \qquad (1 - \sum_{i=1}^{p} a_i L^i) y_t = \varepsilon.$$

Характеристичним поліномом даної AR(p) моделі називають поліном $a(z) = 1 - \sum_{i=1}^{p} a_i z^i$, де z – комплексне число. З курсу загальної алгебри відомо, що такий поліном завжди має p комплексних коренів. Варто відмітити, що цей характеристичний поліном відіграє особливу роль під час перевірки часового ряду на стаціонарність.

Майже у кожному складному часовому ряді присутня так звана випадкова компонента, яку, звичайно, нам теж хотілось би навчитися описувати та аналізувати. Для цього існує так звана модель рухомого середнього (Moving Average) порядку q (MA(q)) часового ряду y_t, яка описується формулою

$$y_t = \varepsilon_t + \sum_{i=1}^{q} b_i \varepsilon_{t-i},$$

де b_i – деякі сталі, ε_{t-1}, ε_{t-2}, …, ε_{t-q} – це значення білого шуму ε в попередні моменти часу. Білий шум ε, як і в моделі AR(p), має $\mathbb{E}(\varepsilon_t) = 0$ та $\mathbb{D}(\varepsilon_t) = \sigma_\varepsilon$.

Якщо використати означення оператора лагу $L^i\varepsilon_t = \varepsilon_{t-i}$, то цю формулу можно переписати у вигляді

$$y_t = (1 + \sum_{i=1}^{q} b_i L^i)\varepsilon_t.$$

Поліном $b(z) = (1 + \sum_{i=1}^{q} b_i z^i)$ називають характеристичним поліномом моделі MA(q), яка описує часовий ряд y_t. Однією з цікавих властивостей моделі MA(q) є те, що якщо всі корені z_j характеристичного полінома $b(z)$ знаходяться в зовнішності одиничного кола, тобто $|z_j| > 1$, то часовий ряд, що описується цим процесом рухомого середнього, є обертовим, тобто його можна представити у вигляді нескінченного процесу авторегресії

$$y_t = c + \sum_{i=1}^{\infty} c_i y_{t-i} + \varepsilon_t,$$

де c та c_i – деякі сталі.

Остання модель, визначення якої нам хотілось би представити в цій роботі, – це модель типу ARMA(p, q). Зрозуміло, що загальний часовий ряд не буде достатньо точно описуватися за допомогою використання лише однієї моделі AR(p) або MA(q), якщо ми хочемо отримати якусь скінчену кількість доданків в цих моделях. Тому на практиці розглядають їх об'єднання. Моделлю ARMA(p, q) часового ряду y_t, де p - кількість регресійних компонент, а q – кількість компонент рухомого середнього, від яких залежить значення y_t початкового часового ряду. За допомогою даних вище означень цей процес можна представити у вигляді

$$y_t = c + \varepsilon_t + \sum_{i=1}^{p} a'_i y_{t-i} + \sum_{i=1}^{q} b'_i \varepsilon_{t-i},$$

де a'_i та b'_i деякі сталі. Якщо переписати через оператор лагу, то отримуємо

$$(1 - \sum_{i=1}^{p} a'_i L^i)y_t = c + (1 + \sum_{i=1}^{q} b'_i L^i)\varepsilon_t.$$

Відмітимо також, що стаціонарний часовий ряд, який описується моделлю ARMA(p, q) можно представити у вигляді нескінченного процесу

рухомого середнього (MA(∞)), а саме

$$y_t = \frac{c}{a(1)} + \sum_{i=0}^{\infty} c_i \varepsilon_{t-i},$$

де c_i – деякі сталі, $a(1)$ – значення характеристичного полінома $a(z) = (1 - \sum_{i=1}^{p} a'_i L^i)$ в точці $z = 1$.

Ця властивість насправді є дуже важливою, оскільки за допомогою неї можна отримати наближення стаціонарного ряду з будь-якою наперед заданою точністю моделлю ARMA(p, q). Пояснення цього факту буде представлено у наступній частині цієї секції.

6.2 Чому так важливо мати стаціонарний ряд?

На прикладах, описаних в секції 5 та секції 4, не важко зрозуміти, що задача прогнозування часових рядів не є тривіальною: в першому випадку нам вдалося дуже точно наблизити моделі прогнозування до початкових часових рядів, проте значення прогнозу співпало зі значенням прогнозу наївної моделі. В секції 4 за допомогою апроксимації часового ряду поліномами вдалося знайти прогноз, відмінний від значень здобутих простими методами, проте залишки $\tau_t = y_t - y'_t$ від значень початкового ряду y_t, де y'_t значення полінома, не розподілені нормально, що, звичайно, свідчить про погану наближеність прогнозуючої моделі до дійсності. Як же досягнути того, щоб побудована нами модель була достатньо близькою до початкового часового ряду і при цьому прогноз був досить високої якості? На це запитання дають відповідь наступні теореми.

Теорема 1. *Кожний стаціонарний часовий ряд можна представити у вигляді рухомого середнього* (Moving Average) *нескінченного порядку* MA(∞).

Теорема 2. *Кожний стаціонарний часовий ряд можно описати моделлю* ARMA(p, q) *з наперед заданою точністю.*

Таким чином, моделі MA(∞) та ARMA(p, q), про які ми будемо ще неодноразово говорити, здатні достатньо точно описати часові ряди. Проте оскільки приведені вище теореми працюють тільки для стаціонарних

рядів, то перше, що нам необхідно зробити – це перевірити часовий ряд „Box office“ на стаціонарність, а у випадку нестаціонарності слід провести деякі перетворення початкового ряду для його стабілізації.

6.3 Теорія, яка лежить в основі тесту Дікі-Фуллера (перевірки ряду на стаціонарність)

Відомо, що існує зв'язок між стаціонарним рядом та його авторегресійною моделлю AR(p). Дійсно, наступна теорема пов'язує стаціонарність (або нестаціонарність) часового ряду з розміщенням на комплексній площині коренів характеристичного полінома $a(z) = 1 - \sum_{i=1}^{p} a_i z^i$ моделі AR(p), що описує даний ряд.

Теорема 3. *Якщо часовий ряд y_1, y_2, ..., y_n описується моделлю AR(p) та всі корені характеристичного полінома належать внутрішності одиничного кола, тобто корені z_i задовольняють нерівності $|z_i| < 1$, то часовий ряд y_1, y_2, ..., y_n є стаціонарним часовим рядом. Якщо існує хоча б один одиничний корінь $|z_i| = 1$, то часовий ряд y_1, y_2, ..., y_n – нестаціонарний.*

Перше, що спадає на думку, – це те, що для перевірки часового ряду на стаціонарність нам потрібно спочатку знайти значення параметрів p та a_i для моделі AR(p), яка найкраще описує наш початковий ряд. Потім знайти всі корені характеристичного полінома та, з'ясувавши їх розміщення на комплексній площині, зробити висновки щодо стаціонарності. Проте існує спосіб обійти цю складну процедуру та перевірити ряд на стаціонарність з використанням приведеної вище теорії.

Насправді, для того, щоб корені z_i характеристичного полінома $a(z)$ задовольняли нерівності $|z_i| < 1$, значення параметрів a_i, звичайно, не може бути довільним. Нерівності, які задовольняють ці значення для відповідних p в моделі AR(p), подані нижче

- для AR(1) необхідно $-1 < a_1 < 1$;
- для AR(2) необхідно $-1 < a_1 < 1$, $a_1 + a_2 < 1$ та $a_2 - a_1 < 1$;

- зі зростанням p обмеження на a_i ускладнюються.

Незалежно від значення p для будь-якої AR(p) існує одне спільне та просте обмеження на перший коефіцієнт моделі авторегресії, а саме: $-1 < a_1 < 1$. Саме тому задача перевірки ряду на стаціонарність зводиться до задачі пошуку коефіцієнту a_1 в авторегресійному рівнянні першого порядку AR(1) для данного часового ряду. Критерій, якій дозволяє статистично перевірити відсутність або наявність одиничного кореня, тобто з'ясувати, чи належить параметр a_1 проміжку $(-1, 1)$, має назву "Критерій Дікі-Фуллера". Саме цей критерій ми і будемо використовувати для перевірки стаціонарності часового ряду „Box office“. Наведемо короткий опис цього критерію.

Критерій Дікі-Фуллера

часовий ряд: $y \quad \Rightarrow \quad y_1,\ y_1, \ldots,\ y_t$;

початкова гіпотеза: $H_0:$ ряд нестаціонарний;

альтернатива: $H_1:$ ряд стаціонарний;

статистика: t-статистика для перевірки значущості

коефіцієнтів лінійної регресії;

розподіл: розподіл Дікі-Фуллера.

Звичайно, в Python є все необхідні бібліотеки для знаходження параметру a_1 в $y_t = a_1 y_{t-1} + \varepsilon_t$ моделі та перевірки ряду на стаціонарність за допомогою критерію Дікі-Фуллера.

6.4 Чи є часовий ряд „Box office“ стаціонарним за критерієм Дікі-Фуллера?

Після підстановки значень часового ряду „Box office“ у відповідну програму, ми отримали наступні значення для ADF (Augmented Dickey–Fuller) Statistic:

```
ADF Statistic: -1.65677
p-value: 0.453512
Critical Values:
                1%: -3.509
                5%: -2.896
                10%: -2.585
```

Отримані результати показують, что часовий ряд „Box office“ не є стаціонарним, оскільки значення ADF Statistic = -1.65677 не виходить за рамки критичних значень (Critical Values) на кожному з трьох рівнів значущості $\alpha \in \{1\%,\ 5\%,\ 10\%\}$.

6.5 Стабілізація дисперсії та знищення трендової компоненти

Згідно з теоремою про наближення стаціонарного часового ряду (Теорема 2), використати модель типу ARMA(p, q) для часового ряду „Box office“ ми зможемо тільки у тому випадку, якщо за допомогою скінченої кількості перетворень нам вдасться зробити його стаціонарним. Після того, як прогноз для стаціонарного ряду буде зроблений, потрібно повернутися до початкових змінних за допомогою обернених перетворень.

Першою причиною нестаціонарності часового ряду „Box office“ є зростання дисперсії ряду з плином часу. Дійсно, на проміжку (t_1, t_{30}) дисперсія дорівнює 5235.41, на проміжку (t_{31}, t_{60}) – 5861.16, а на проміжку (t_{61}, t_{90}) вже дорівнює 164746.42, що суттєво відрізняється від всіх інших значень. На практиці дисперсію стабілізують за допомогою сім'ї перетворень Бокса-Кокса, які описуються формулою:

$$s_t = \begin{cases} \ln y_t & , \lambda = 0, \\ (y_t^\lambda - 1)/\lambda & , \lambda \neq 0, \end{cases}$$

де параметр λ підбирають таким чином, щоб розподіл результуючої послідовності елементів часового ряду s_t був максимально наближеним до нормального розподілу. Одним зі способів відшукання оптимального значення цього параметру є максимізація логарифма функції правдоподі-

бності:

$$f(x,\lambda) = -\frac{N}{2}\ln\left[\sum_{i=1}^{N}\frac{\left(s_i(\lambda)-\bar{s}(\lambda)\right)^2}{N}\right] + (\lambda-1)\sum_{i=1}^{N}\ln(y_i),$$

де

$$\bar{s}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} s_i(\lambda).$$

Більш детально ми не будемо зупинятися на описі способів оптимізації параметра λ.

За допомогою відповідних програм було отримано, що для часового ряду „Box office“ потрібно робити перетворення з параметром $\lambda = 0.09833$. Таким чином, після застосування перетвореня Бокса-Кокса до ряду y_t, отримуємо перетворений часовий ряд s_t, графік якого представлений нижче.

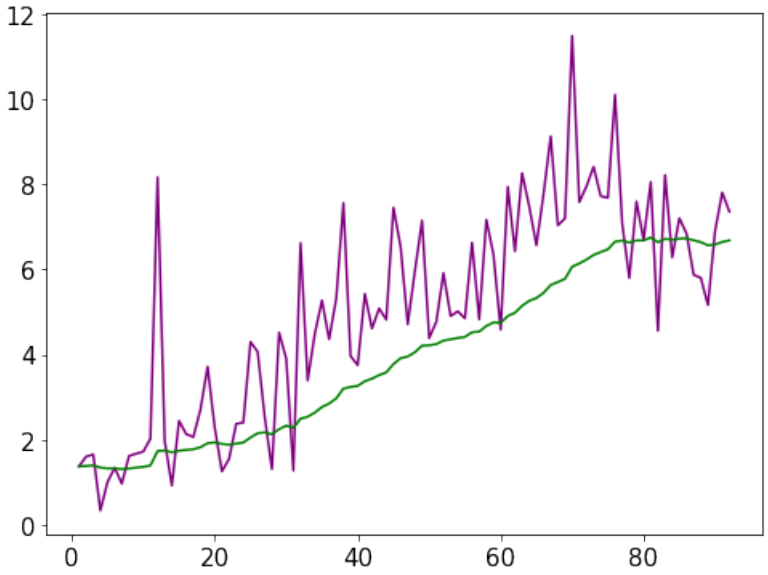

Рис. 12: Часовий ряд „Box office“ пісня застосування перетворення Бокса-Кокса

Для перевірки отриманого ряду на стаціонарність знову застосуємо критерій Дікі-Фуллера.

Отриманий часовий ряд s_t, як і початковий, ще не є стаціонарним рядом, проте вже причина нестаціонарності полягає у наявності трендової компоненти (зелена лінія), яка, звичайно, змінює значення середнього для фіксованого інтервалу часу. Дійсно, для інтервалу (t_1, t_{30})

```
ADF Statistic: -1.756486
p-value: 0.402269
Critical Values:
                1%: -3.510
                5%: -2.896
                10%: -2.585
```

середнє арифметичне дорівнює 2.32, для (t_{31}, t_{60}) – 5.23 та для (t_{61}, t_{90}) – 7.364807654. Щоб позбавитися від трендової компоненти іноді достатньо початкового ряду перейти до ряду його папарних різницях (диференціювання часового ряду). Зазначимо, що диференціювання ряду можно робити декілька разів поки ряд не стане стаціонарним. Необхідне перетворення $s_1,\ s_2,\ \ldots,\ s_t \to s'_2,\ s'_3,\ \ldots,\ s'_t$ здійснюється за допомогою формули

$$s'_t = s_t - s_{t-1}.$$

При цьому, кількість значень часового ряду зменшується, проте вже після першого диференціювання початкова гипотеза H_0 в тесті Дікі-Фуллера відхиляється навіть при $\alpha = 1\%$ рівні значущості, що говорить про стаціонарність ряду s'_t. Таким чином ми перетворили початковий

```
ADF Statistic: -6.195013
p-value: 0.000000
Critical Values:
                1%: -3.510
                5%: -2.896
                10%: -2.585
```

нестаціонарний ряд y_t „Box office“ на стаціонарний ряд s'_t. Само для цього ряду ми і будемо шукати оптимальну модель ARMA(p,q) та робити прогноз на декілька років вперед, а потім, за допомогою обернених перетворень, знайдемо значення прогнозу для початкового ряду y_t. Графік стаціонарного ряду s'_t поданий нижче.

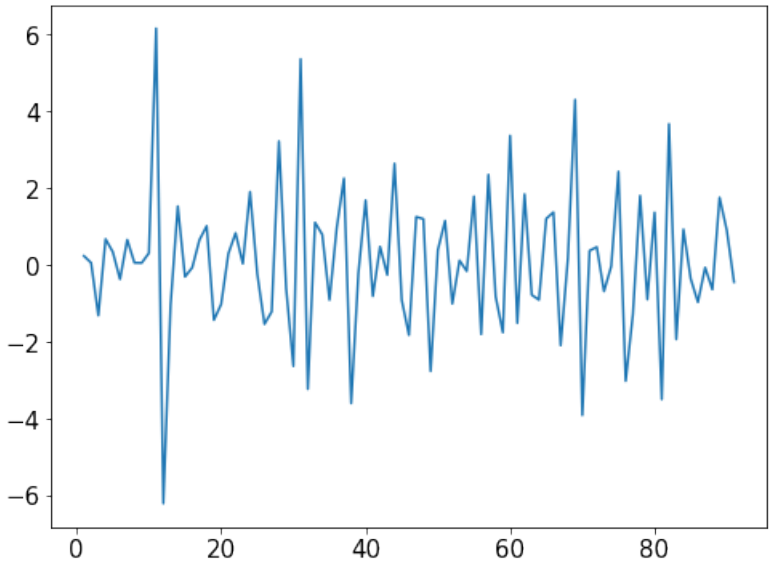

Рис. 13: Стаціонарний ряд, отриманий з часового ряду „Box office“

6.6 Як знайти коефіцієнти p та q в моделі ARMA(p,q)

Нагадаємо, що процес ARMA(p,q), якій ми розглядаємо для стаціонарного ряду s_t в цій секції описується формулою

$$s'_t = c + \varepsilon_t + \sum_{i=1}^{p} a'_i y_{t-i} + \sum_{i=1}^{q} b'_i \varepsilon_{t-i},$$

де c, a'_i та b'_i деякі сталі. Очевидним є те, що для знаходження прогнозу за допомогою такої моделі нам спочатку потрібно з'ясувати, яку кількість регресійних компонент p та яку кількість компонент рухомого середнього q нам необхідно розглядати, щоб дана модель ARMA(p,q) мала найменші відхилення від початкового стаціонарного ряду s'_t. Зрозуміло, якщо ми будемо брати довільні значення цих параметрів, то знайдений прогноз може бути далеким від дійсності.

Почнемо, з аналізу коефіцієнтів b'_i рухомого середнього. Насправді вони дуже погано піддаються інтерпретації, тому складно стверджувати, яке саме значення параметру q слід використовувати в моделі ARMA(p,q). Проте існує функція, яка дозволяє оцінити вплив доданків ε_{t-1}, ε_{t-2}, ... на значення часового ряду s'_t. Це здійснюється за допомогою автокореляційної функції, означення якої ми вже наводили у підсекції «Чи потрібно позбавлятися від викидів?». Нагадаємо більш стисло,

що автокореляцією порядку k називають кореляцію між часовим рядом s'_t та його копією зсунутою на k значень, тобто

$$r_k = Corr(y_t, y_{t-k}).$$

Функцією автокореляції називають залежність значення коефіцієнту r_k від значення відповідного лагу k. Для стаціонарного процесу r_k показують на скільки в середньому зміниться значення часового ряду s'_t, при умові зміни середнього значення часового ряду в точці s'_{t-k} на одиницю[2]. Припустимо, що всі коефіцієнти автокореляції, починаючи з r_q, дорівнюють нулю. Для стаціонарного процесу це фактично значить, що всі значення після s'_{t-q} припиняють в середньому діяти на значення часового ряду в точці s'_t. Проте, як відомо, в моделі ARMA(p,q) значення часових рядів залежать не тільки від регресійних компонент, а ще й від компонент рухомого середнього. Тому останнє значення q, при якому r_q не дорівнює нулю, можна інтерпретувати, як останнє значення при якому регресійні компоненти та компоненти рухомого середнього одночасно припиняють впливати на значення часового ряду s'_t. Загальна процедура знаходження коефіцієнту q в моделі ARMA(p,q) полягає лише в побудові функцією автокореляції для стаціонарного часового ряду s'_t. **Останнє значення k при якому r_k суттєво відрізняється від нуля і буде нашим шуканим значенням**.

Принцип відшукання значення параметру p в моделі ARMA(p,q) будується аналогічним чином, проте його оцінка здійснюється вже не за допомогою функції автокореляції, а за допомогою часткової функції автокореляції. Часткова автокореляція порядку k визначається за формулою

$$\phi_k = Cor(s'_t - P(s'_t), s'_{t-k} - P(s'_{t-k})),$$

де $P(s'_t)$ – проекція випадкової величини s'_t на лінійну оболонку значень $s'_{t-1}, s'_{t-2}, \ldots, s'_{t-k+1}$.

Ідея знаходження цього коефіцієнту полягає у мотивації знайти вплив значення стаціонарного часового ряду в точці s'_{t-k} на значення відповід-

[2]Така інтерпретація значення коефіцієнту r_k є тільки для стаціонарного часового ряду.

ного часового ряду в точці s'_t, проте вже без впливу всіх інших значень s'_{t-1}, s'_{t-2}, ..., s'_{t-k+1}. В цьому і полягає основна відмінність часткової автокореляції від звичайної автокореляції, оскільки в автокореляції k-го порядку на зміну значення s'_t впливали зміни не лише в точці s'_{t-k}, а ще і у всіх інших точках часового ряду s'_{t-1}, s'_{t-2}, ..., s'_{t-k+1}. Припустимо, що всі значення часткової автокореляції після ϕ_k починають дорівнювати нулю. Для стаціонарного часового ряду s_t, який описується моделлю ARMA(p,q), фактично це свідчить про те, що на значення часового ряду в точці s'_t вже не впливають інші регресійні компоненти після s'_{t-p}. Таким чином, загальна процедура знаходження коефіцієнту p в моделі ARMA(p,q) полягає лише в побудові функції часткової автокореляції для стаціонарного часового ряду s'_t. **Останнє значення k при якому ϕ_k суттєво відрізняється від нуля і буде нашим шуканим значенням**.

6.7 Значення коефіцієнтів p та q в моделі ARMA(p,q) для часового ряду „Box office“

Використовуючи процедуру, про яку ми говорили в попередній підсекції, знайдемо значення коефіцієнтів p та q в моделі ARMA(p,q) для стаціонарного ряду s'_t, отриманного за допомогою перетворення Бокса-Кокса та першого диференціювання часового ряду „Box office“.

Функції автокореляції та часткової автокореляції зображені на графіках **Рис.14** та **Рис.15** відповідно.

На графіку функції автокореляції можна побачити, що тільки автокореляції 1-го порядку суттєво відрізняється від нуля, а на графіку часткової автокореляції відміною від нуля будемо вважати тільки часткову автокореляцію 2-го порядку. Таким чином ми з'ясували, що оптимальную моделлю для прогнозування стаціонарного часового ряду є модель ARMA(2,1). Саме її ми будемо застосовувати для подальшого аналізу.

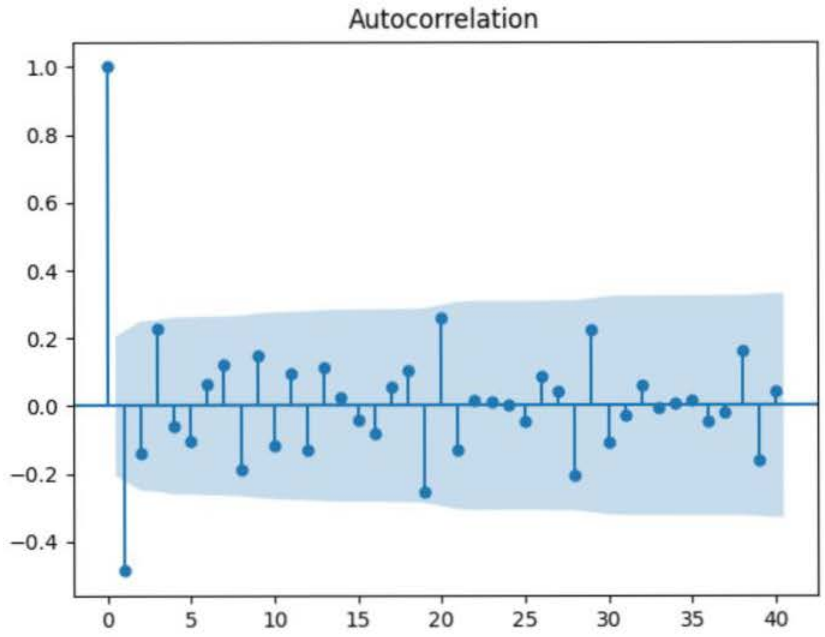

Рис. 14: Графік функції автокореляції для стаціонарного ряду s'_t

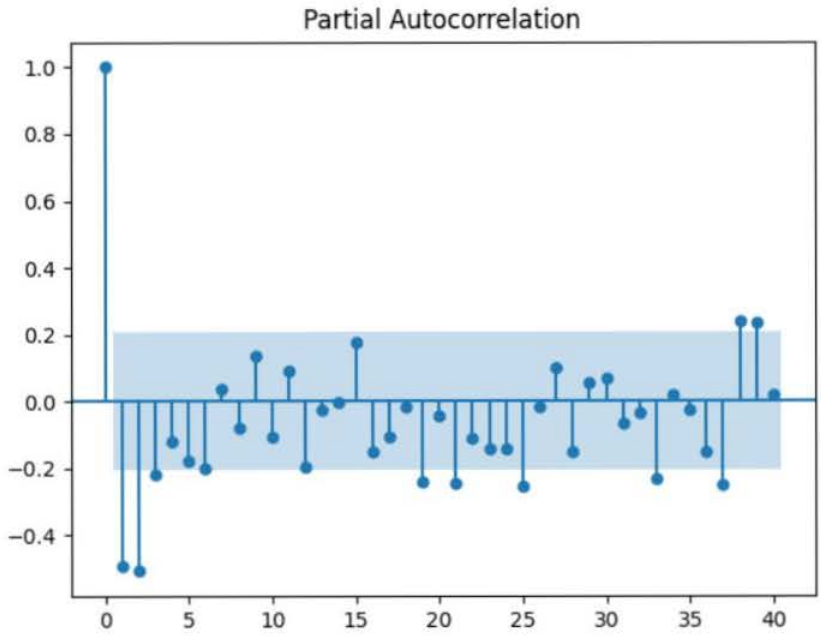

Рис. 15: Графік функції часткової автокореляції для стаціонарного ряду s'_t

6.8 Прогнози ARMA(2,1)

Використаємо функцію ARMA в Python для підбору коефіцієнту c, коефіцієнтів a'_1 та a'_2 регресійної компоненти та коефіцієнта b'_1 в компоненті рухомого середнього прогнозуючої моделі ARMA(2,1) для стаціонарного

ряду s'_t. Нагадаємо, що s'_t – стаціонарний ряд, отриманий з часового ряду „Box office“ за допомогою перетворення Бокса-Кокса та першого диференціювання. Після підстановки відповідних параметрів $p = 2$ та $q = 1$ отримуємо відповідний опис моделі в Python **Рис.16**.

```
                              ARMA Model Results
==============================================================================
Dep. Variable:            First diff.   No. Observations:                   91
Model:                 ARMA(2, 1)   Log Likelihood                -160.536
Method:                    css-mle   S.D. of innovations              1.403
Date:                Sat, 22 May 2021   AIC                            331.071
Time:                        11:44:24   BIC                            343.626
Sample:                             0   HQIC                           336.136

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0638      0.038      1.681      0.093      -0.011       0.138
ar.L1.First diff.    -0.2833      0.155     -1.833      0.067      -0.586       0.020
ar.L2.First diff.    -0.2610      0.132     -1.982      0.048      -0.519      -0.003
ma.L1.First diff.    -0.6108      0.142     -4.294      0.000      -0.890      -0.332
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.5428           -1.8807j            1.9575           -0.2947
AR.2           -0.5428           +1.8807j            1.9575            0.2947
MA.1            1.6372           +0.0000j            1.6372            0.0000
-----------------------------------------------------------------------------
```

Рис. 16: Опис моделі ARMA(2,1) в Python

Значення термінів, що представлені на **Рис.16** ми пояснювати не будемо. Нас цікавить лише значення стовпців **coef** та **std err**, в яких містяться значення шуканих коефіцієнтів та їх середнє квадратичне відхилення відповідно. Графік прогнозу стаціонарного ряду s'_t зображений на малюнку **Рис.17**. Окрім, часового ряду s'_t та його прогнозу на цьому графіку є 95% довірчий інтервал для майбутніх значень. Одразу постає питання, чому довірчий інтервал для прогнозу вийшов настільки широким. Відповідь на це представлена нижче, при аналізі залишків $r_t = y_t - \hat{y}_t$, де $\hat{y}_t$ – значення прогнозу моделі $ARMA(2, 1)$, від значень початкового часового ряду y_t „Box office“. Щоб знайти прогноз початкового часового ряду „Box office“ нам потрібно зробити два обернених перетворення. Графік прогнозу оберненого перетворення для першого диференціювання зображений на **Рис.18**. Залишилось тільки знайти обернене до перетворення Бокса-Кокса, щоб отримати прогноз часового ряду „Box office“ за допо-

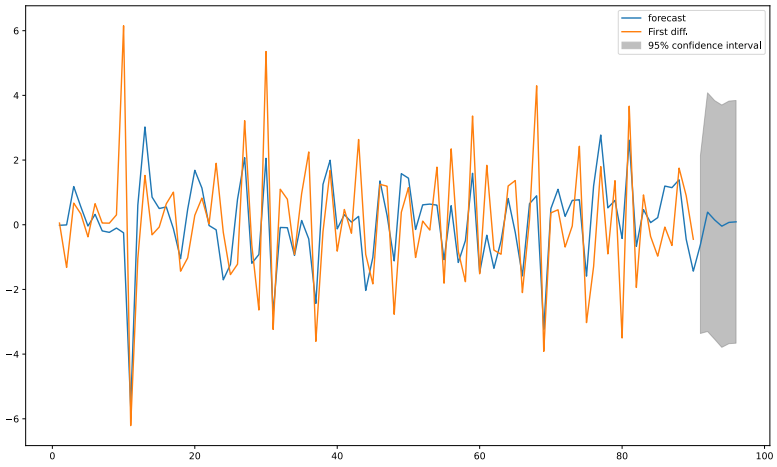

Рис. 17: Прогноз моделі ARMA(2,1) для стаціонарного ряду s'_t

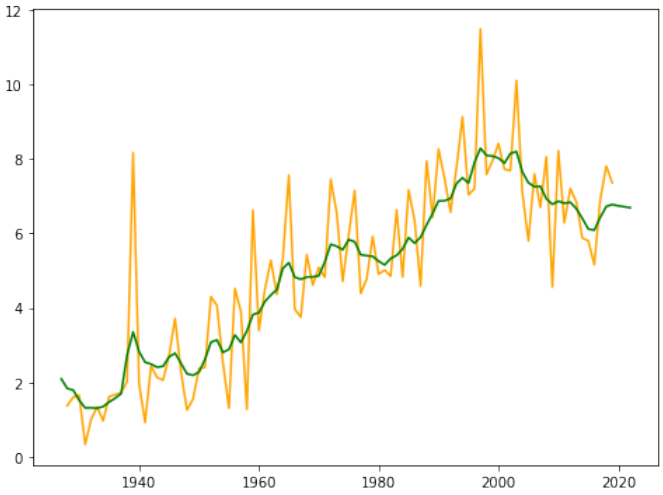

Рис. 18: Прогноз моделі ARMA(2,1) для ряду s_t

могою моделі ARMA(2,1). Кінцевий результат зображений на малюнку **Рис.19**.

Перевіримо залишки $r_t = y_t - \hat{y}_t$ від прогнозу на нормальність, де y_t – значення початкового ряду, $\hat{y}_t$ – значення прогнозу моделі $ARMA(2,1)$ для початкового часового ряду „Box office“. На графіках **Рис.20** та **Рис.21** зображені гістограма значень залишків r_t та графік Q-Q для значень залишків r_t.

Як бачимо, гістограма розподілу **Рис.20** несиметрична, а точки на графіку Q-Q **Рис.21** не лежать на одній прямій, тому ми можемо при-

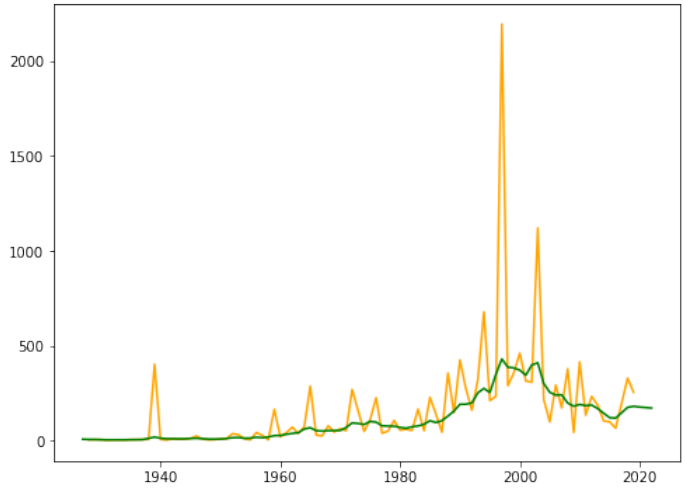

Рис. 19: Прогноз моделі ARMA(2,1) для часового ряду „Box office“

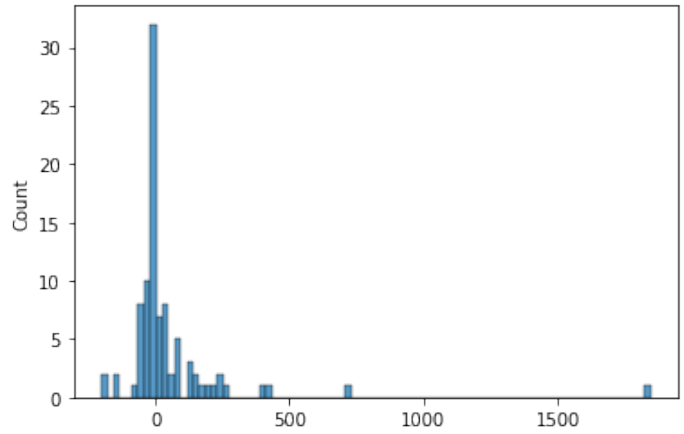

Рис. 20: Гістограма значень залишків r_t

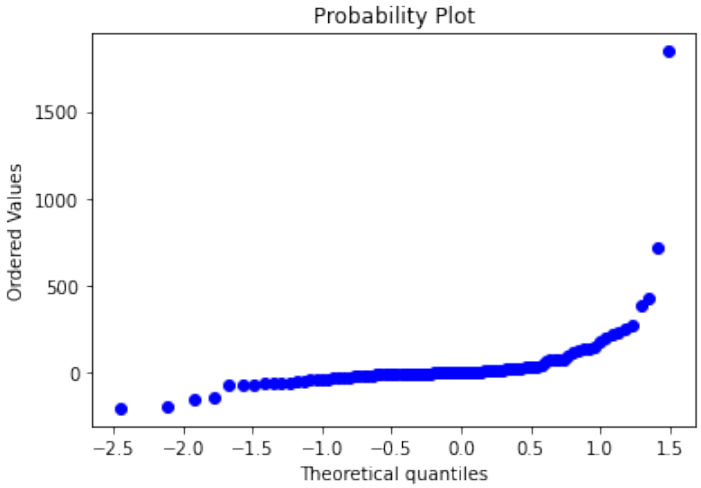

Рис. 21: Графік Q-Q для значень залишків r_t

пустити, що розподіл залишків не є нормальним. На графіку Q-Q **Рис.21** можна помітити, що, порівняно з нормальним розподілом, в центрі розподілу знаходиться набагато більше даних, ніж на краях. Про такий розподіл говорять, що він має важкий хвіст (heavy-tailed) **Рис.22**.

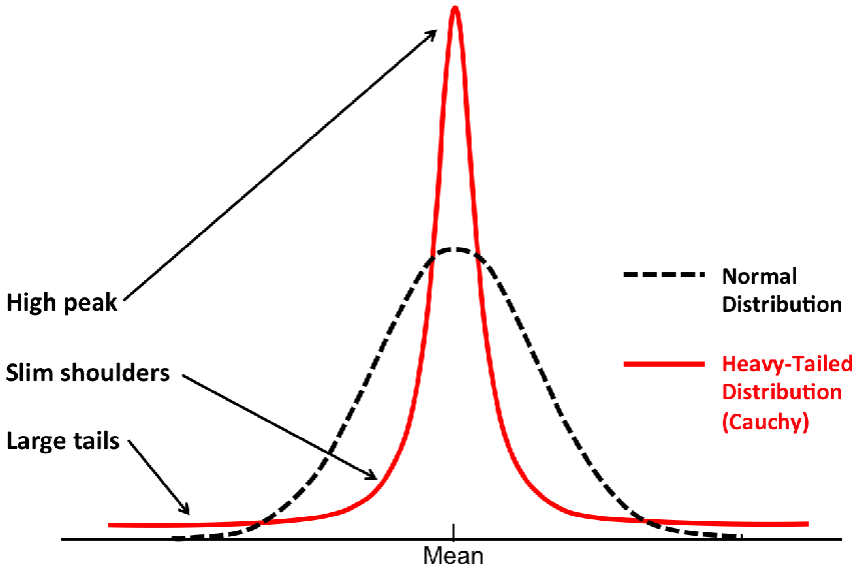

Рис. 22: Важкий хвіст

З загальної теорії про ARMA-процеси відомо, що довірчі інтервали для прогнозу цієї моделі будуються в залежності від залишків. Якщо подивитись на **Рис.17**, то можна побачити, що 95% довірчий інтервал вийшов несподівано великим. Це зумовлено тим, що розподіл залишків r_t має важкий хвіст, тому квантиль рівня 95% знаходиться дуже далеко від середнього значення розподілу.

Незважаючи на всі недоліки підібраної ARMA-моделі, ми вважаємо за потрібне виписати значення прогнозу ряду „Box office“ на наступні 4 роки:

$$y_{2021} = 180,2102832$$

$$y_{2022} = 176,2714942$$

$$y_{2023} = 173,7009842$$

$$y_{2024} = 171,0292946$$

7 Висновки

Єдиний метод прогнозування, за допомогою якого був отриманий фільм-переможець – це логістична регресія. За її прогнозом у 2021 році мав перемогти фільм „The Trial of the Chicago 7“. Хоча і переміг інший фільм („Nomadland“), але сайти та статті критиків ставили спрогнозований нами фільм на друге місце. Можливо, якби ми враховували у цій моделі не тільки рейтинги фільмів, а й інші параметри, то наш прогноз міг би бути більш вдалим.

Всі інші методи прогнозування, які базуються на часових рядах, виявились незастосовними для прогнозування такої складної системи, тому були використані та описані в цій роботі з навчальною метою. Насправді, для якісного прогнозу треба було б зібрати набагато більше видів даних та дослідити особисті уподобання членів жюрі, що обирають переможця. Оскільки тут присутній людський фактор, то математичний опис цієї системи є дуже складною задачею.